

Practice of Epidemiology

Gene-Environment Interaction in Genome-Wide Association Studies

Cassandra E. Murcay, Juan Pablo Lewinger, and W. James Gauderman

Initially submitted November 27, 2007; accepted for publication May 5, 2008.

It is a commonly held belief that most complex diseases (e.g., diabetes, asthma, cancer) are affected in part by interactions between genes and environmental factors. However, investigators conducting genome-wide association studies typically test for only the marginal effects of each genetic marker on disease. In this paper, the authors propose an efficient and easily implemented 2-step analysis of genome-wide association study data aimed at identifying genes involved in a gene-environment interaction. The procedure complements screening for marginal genetic effects and thus has the potential to uncover new genetic signals that have not been identified previously.

association; environment; genes; genetic markers; genetics; genome

Abbreviations: E, environment; G, gene; GWAS, genome-wide association study; OR, odds ratio; SNP, single nucleotide polymorphism.

Editor's note: Two invited commentaries on this article appear on pages 227 and 231, and the authors' response is published on page 234.

Many common, complex traits are believed to be a result of the combined effect of genes, environmental factors, and their interactions. For example, Ito et al. (1) showed a significant interaction between smoking status and the apurinic/apyrimidinic endonuclease 1 protein coding gene (*APE1*) for lung cancer. Stern et al. (2) found smoking status to be an effect modifier of the association between the XPD codon 751 polymorphism and risk of bladder cancer. Understanding the relation between genetic polymorphisms and environmental exposures can help to identify high-risk subgroups in the population and provide better insight into pathway mechanisms for complex diseases.

Methods for identifying disease susceptibility genes include linkage analysis, candidate gene association studies, and, more recently, the genome-wide association study (GWAS). It is known that a GWAS can be more powerful than linkage analysis in detecting genes associated with modest increases in disease risk (3). Current GWAS meth-

ods are designed to detect main effects, that is, direct associations of a single nucleotide polymorphism (SNP) or clusters of SNPs with disease (4, 5). In the context of complex diseases, scanning for main effects might miss important genetic variants specific to subgroups of the population, as defined by some exposure. In fact, interactions with opposite effects in 2 different exposure groups (crossing-interaction) will not show a main effect and therefore will not be identified by using standard approaches.

Despite the importance of gene-environment interaction for complex diseases, little work has been done to develop methods for detecting these types of interactions in the context of a GWAS. In this paper, we focus on identifying SNPs that demonstrate heterogeneity between subgroups defined by some environmental exposure. We introduce an efficient 2-step approach for detecting loci involved in gene-environment interactions that is performed independently of any initial scans for main effects. Our method expands on the traditional test for gene-environment interaction in a case-control study by incorporating a preliminary screening step constructed to efficiently use all available information in the data. We demonstrate that this 2-step approach is more powerful than the standard test of interaction across a wide range of models.

MATERIALS AND METHODS

Let D be an indicator of disease status, and assume we have a sample of cases ($D = 1$) and unrelated controls ($D = 0$). Assume that information is available for a binary environmental exposure, with E as an indicator for exposure. Further assume that for each individual we have genotyped M SNPs spanning the genome, with g_1, g_2, \dots, g_M denoting the genotypes at the M loci. Letting G_1, G_2, \dots, G_M denote some genetic coding (e.g., additive, dominant) for each genotype, we consider a model for a given SNP of the form

$$\text{logit } P(D = 1 | g, e) = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} GE. \quad (1)$$

Under a dominant coding of the genotype, for example, $\exp(\beta_g) = \text{OR}_g$ is the odds ratio (OR) comparing carriers of at least 1 risk allele ($G = 1$) with noncarriers ($G = 0$) among those unexposed ($E = 0$). $\exp(\beta_e) = \text{OR}_e$ is the odds ratio comparing risk for exposed ($E = 1$) with that for unexposed ($E = 0$) individuals among noncarriers of the risk allele. Lastly, $\exp(\beta_{ge}) = \text{OR}_{ge}$ is the ratio of the genetic odds ratios comparing exposed with unexposed subjects, that is, $\text{OR}_{g|E=1} / \text{OR}_{g|E=0}$. If this ratio is equal to 1.0, or $\beta_{ge} = 0$, we say that there is no interaction between genotype and the environmental exposure.

In the context of a GWAS, a standard approach to test for $G \times E$ interaction would be to perform a 1-df test of $H_0 : \beta_{ge} = 0$ for each SNP based on the model in equation 1. We assume that a likelihood ratio test will be used to test this hypothesis and denote it as the “1-step” test of interaction. A correction for multiple comparisons (e.g., Bonferroni, controlling the false discovery rate (6)) is required to achieve a desired genome-wide type I error.

One might consider using a case-only analysis of diseased individuals as the sole approach for identifying interactions in a GWAS. Indeed, the case-only test has been shown to be more powerful than a case-control analysis for identifying interactions (7, 8). However, the case-only analysis depends heavily on an underlying assumption of G-E independence in the population, which would be untenable across all SNPs being scanned in a GWAS. If there is an underlying population association between genotype and environmental exposure, a case-only analysis will result in an inflated number of false positives (7, 9).

We propose an alternative 2-step test to scan for interactions that combines the power of the case-only test with the protection from bias of the case-control test. Our 2-step scan for interactions consists of the following 2 steps:

1. Step 1, screening test: For each of the M SNPs, we perform a likelihood ratio test of association between G and E based on the logistic model $\text{logit } P(E = 1 | g) = \gamma_0 + \gamma_g G$. This is the standard test that would be applied in a case-only analysis of $G \times E$ interaction (7, 9), although, in our context, the test is applied to the combined sample of cases and controls. The subset of m SNPs that exceeds a given significance threshold (i.e., with $P < \alpha_1$) for the test of $H_0 : \gamma_g = 0$ is analyzed in step 2.
2. Step 2, case-control test: The m SNPs that pass step 1 are assessed in the traditional test of $G \times E$ interaction, that

is, based on a likelihood ratio test of $H_0 : \beta_{ge} = 0$ derived from the model in equation 1. Significance at this step is defined as having a P value less than α/m , where α is the desired overall type I error rate.

Although step 1 of our procedure is also sensitive to the assumption of independence between G and E in the population, the step 2 comparison of cases to controls is not. Therefore, our overall 2-step procedure will provide a valid test in the presence of population-level association between genotype and exposure, a claim we will verify by simulation.

Given the reported power of a case-only analysis applied to only diseased individuals (7, 9), one may be tempted to apply step 1 to only diseased individuals, that is, to perform a true case-only test of interaction in step 1 and use it to define the subset of m markers to analyze in step 2. However, this approach produces a correlation between the step 1 and step 2 test statistics and leads to an inflated type I error rate for the overall procedure. Our screening test of $G \times E$ interaction applied to the entire sample of cases and controls eliminates the correlation between tests in steps 1 and 2 (Appendix 1) and, as we show (Appendix 2) and verify by simulation, preserves the overall type I error rate.

We performed simulations to study the power achieved by our 2-step testing framework compared with the traditional 1-step method. For each of 1,000 replicate data sets, we simulated a sample of 500 cases and 500 controls, each with genotype information on a large number of markers ($M = 10,000, 25,000$, and $50,000$). Although larger marker sets are likely to be used in practice, our chosen set sizes are sufficient to demonstrate the relative power of our 2-step method. Markers were assumed to be independent loci distributed across the genome. For each replicate, a single marker was chosen to be the true disease susceptibility locus, with remaining markers assumed to have no association with disease. We considered a range of minor allele frequencies (q_A) for the disease susceptibility locus, including 0.10, 0.20, and 0.30. For the remaining null markers, we simulated a uniform distribution of allele frequencies between 0.05 and 0.30. We also considered a range of values for the exposure prevalence (p_E), including 0.10, 0.25, and 0.50, and set the population disease prevalence to 0.05. Finally, we considered a range of possible values for the genetic and environmental main effects (β_g and β_e , respectively) as well as for the interaction effect (β_{ge}).

As described above, the traditional case-only method of testing for $G \times E$ interaction is based on the assumption of no population-level association between the G and E . We explored the sensitivity of our method to population-level association between G and E by introducing a parameter p_{ge} , defined to be the probability that a given marker is associated with the exposure at the population level. With $p_{ge} = 0.0$, none of the markers was assumed to be associated with E in the population. It is unlikely that a large percentage of SNPs will be associated with a given environmental factor, but, for completeness, we considered a wide range of values—0.01 to 0.95. For each simulated marker, we randomly decided whether it was associated with E in the population based on probability p_{ge} . For any marker

chosen to have an association, we generated genotypes conditional on the assigned E and an assumed population marker–exposure odds ratio of 2.0.

For each parameter setting, we applied the traditional 1-step $G \times E$ test and our 2-step approach. We estimated the experiment-wise type I error as the proportion of 1,000 replicates in which at least one of the null markers was found to be significant after a Bonferroni correction for multiple comparisons. Power was calculated as the number of replicates in which the disease susceptibility locus was detected at an overall significance level of 0.05, again after a Bonferroni correction for multiple comparisons. We also computed the proportion of replicates in which the disease susceptibility locus was among the top 10 or top 25 most significant SNPs, that is, on a short list that might warrant additional scrutiny following a genome-wide scan.

RESULTS

Across a range of interaction effect sizes ($R_{ge} = \exp(\beta_{ge})$), our 2-step method was more powerful than the standard 1-step test for detecting an interaction (Figure 1). For example, when $R_{ge} = 3.0$, power was 33.2% using a standard 1-step approach compared with 57.9% using our 2-step method. As we would expect, as the effect size of the interaction increases, both tests gain power. For a small interaction effect, both tests have low power to detect a causal locus; at a sufficiently large effect size, the 2 tests approach 100% power. The largest differences in power between the 2 methods occurred when the interaction effect was of moderate magnitude, from 2.5 to 4.0. The estimates of power in Figure 1 all assumed a disease susceptibility locus allele frequency $q_A = 0.2$, exposure frequency $p_E = 0.5$, no main effects ($R_g = R_e = 1$), no population-level association between g and E ($p_{ge} = 0$), 10,000 SNP markers, and a first-step significance threshold of $\alpha_1 = 0.05$.

For various alternatives to the above parameter settings, Table 1 shows type I error and power for the 1- and 2-step methods for detecting G - E interaction, holding the interaction effect size fixed ($R_{ge} = 3.0$). Both methods approximated the nominal 0.05 type I error under all scenarios, even when there was a population-level association ($p_{ge} \neq 0$) between markers and the environmental factor (also refer to the Web Figure, which is posted on the *Journal's* website (<http://aje.oupjournals.org/>)). The mean type I error across all scenarios was slightly smaller for the 2-step test (mean = 0.051) than for the conventional 1-step test (mean = 0.056), indicating that, on average, the 2-step test was slightly more conservative than the 1-step test. The 2-step test was consistently more powerful than the traditional 1-step case-control test over a wide variety of parameter settings. As expected, power for both tests was highest for common exposures and alleles. The 2-step method was at least twice as powerful as the 1-step test when the exposure was rare or when the disease allele was rare, although absolute power in these situations was low for both procedures. Power for the 2-step test depended somewhat on the significance threshold for step 1 (α_1). Specifically, relative to our base model with $\alpha_1 = 0.05$, a smaller threshold value ($\alpha_1 = 0.01$) resulted in increased power to detect the disease susceptibility locus,

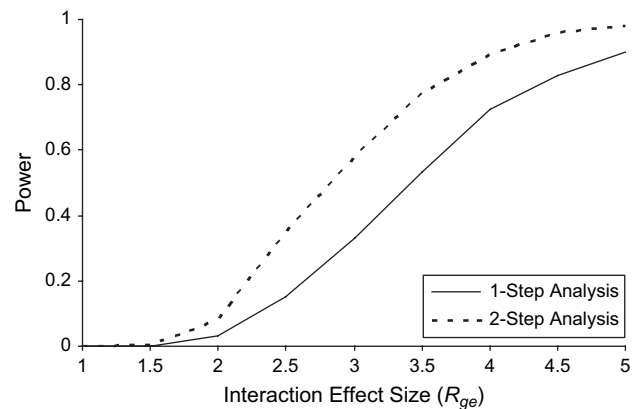


Figure 1. Power for 1-step and 2-step analyses for varying levels of interaction effect size (R_{ge}). All other parameter settings remain constant under “base” model specifications ($M = 10,000$, number of cases/controls = 500/500, $q_A = 0.2$, $p_E = 0.5$, $R_g = 1$, $R_e = 1$, $p_{ge} = 0$, $\alpha_1 = 0.05$).

while we saw reduced power when we allowed more markers to move into step 2 ($\alpha_1 = 0.10$).

As expected, a population-level association between markers and environment ($p_{ge} > 0$) increased the number of markers that proceeded to step 2. However, for the range of values we considered plausible in a genome-wide scan ($p_{ge} = 0.01$ or 0.05), there was not an appreciable impact on power using the 2-step method. At more liberal values for the proportion of markers with a population-level association between G and E ($p_{ge} = 0.30, 0.95$), power for the 2-step method approached that for the traditional 1-step method. Specifically, when we assumed that 95% of the markers were associated with the environmental factor, power estimates for the 1- and 2-step methods were identical (29.3%). Power estimates under the full range of values for p_{ge} we considered are shown in the Web Figure.

Across a range of interaction effect sizes ($R_{ge} = \exp(\beta_{ge})$), our 2-step method was also more likely than the 1-step method to detect the disease susceptibility locus according to the ranked P -value statistics (Figure 2). Using the ordered P values compared with a traditional significance threshold to choose a subset of highly ranked SNPs did not depend as much on interaction effect size to be confident that the selected subset included the disease susceptibility locus. For example, our 2-step approach resulted in the disease susceptibility locus being in the top-10-ranked P values in 79% of the 1,000 replicates with an interaction effect size of 2.5 compared with 59% when the 1-step method was used. Power for this interaction effect size was less than 35% for both tests (Figure 1).

Our 2-step method was more robust to changes in exposure prevalence, minor allele frequency, number of markers, and so forth, when we focused on the rank statistics (Table 2). Under our base model settings, the disease susceptibility locus was in the top 10 P values in 94% of the 1,000 replicates using the 2-step method compared with only 79% of the replicates when the 1-step method was used. Even under the least ideal circumstances, with a low exposure

Table 1. Type I Error and Power for 1-Step and 2-Step Tests for Gene-Environment Interaction

Model ^a	Type I Error ^b		Power ^c	
	1 Step	2 Step	1 Step	2 Step
Base ^d	0.062	0.045	0.332	0.579
Disease susceptibility locus allele frequency (q_A)				
0.1	0.062	0.043	0.158	0.348
0.3	0.052	0.039	0.344	0.588
Exposure prevalence (p_E)				
0.1	0.058	0.073	0.035	0.111
0.25	0.053	0.052	0.212	0.448
Effect sizes (R_g, R_e, R_{ge})				
123	0.063	0.038	0.228	0.479
213	0.054	0.054	0.312	0.578
223	0.053	0.054	0.253	0.483
Population gene-environment association (p_{ge})				
0.01	0.054	0.040	0.320	0.564
0.05	0.057	0.050	0.316	0.504
0.30	0.061	0.049	0.315	0.385
0.95	0.056	0.052	0.293	0.293
No. of markers (M)				
25,000	0.051	0.063	0.290	0.518
50,000	0.049	0.060	0.223	0.436
Step 1 threshold (α_1)				
0.01	0.065	0.049	0.350	0.698
0.1	0.051	0.048	0.330	0.500

^a Variations to model specification refer to a change to a parameter setting in the "base" model. All other parameters remain constant.

^b All estimates of type I error have a standard error of <0.008.

^c All estimates of power have a standard error of <0.016.

^d $M = 10,000$, number of cases/controls = 500/500, $q_A = 0.2$, $p_e = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_{ge} = 0$, $\alpha_1 = 0.05$.

prevalence, the ranked P value for the disease susceptibility locus was still in the top 25 in 63% of replicates using the 2-step method compared with 35% when the 1-step method was used. Under all other scenarios we examined, the ranked P value using the 2-step method was in the top 25 for at least 85% of replicates, and the 2-step method always outperformed the 1-step approach.

DISCUSSION

For a GWAS in a case-control sample, we have shown that our 2-step testing approach provides a powerful alternative for testing $G \times E$ interaction relative to a traditional 1-step test. For the parameter settings we examined, the 2-step method was always more powerful than the 1-step method. Our method was also more robust than the traditional case-control test to changes in allele frequency, exposure prevalence, and other parameters when comparing the ranked

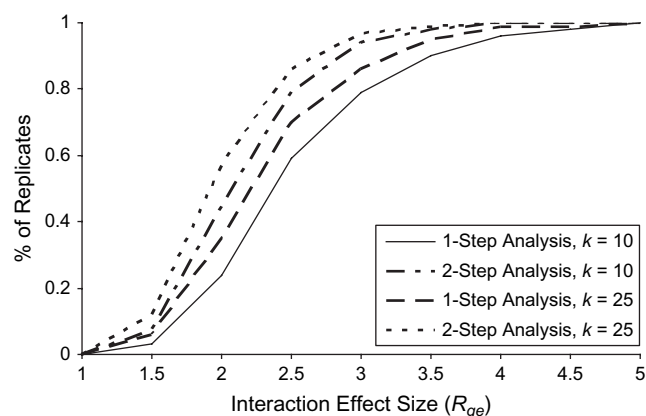


Figure 2. Percentage of replicates for which the P value for disease susceptibility locus is ranked in the top k ($k = 10$ or $k = 25$) marker P values for varying levels of interaction effect size (R_{ge}). All other parameter settings remain constant under "base" model specifications ($M = 10,000$, number of cases/controls = 500/500, $q_A = 0.2$, $p_e = 0.5$, $R_g = 1$, $R_e = 1$, $p_{ge} = 0$, $\alpha_1 = 0.05$).

P value for the true disease-susceptibility locus. Given its increased power and ease of implementation, the 2-step test is an attractive alternative for identifying $G \times E$ interactions in a GWAS for complex diseases.

We assumed a dichotomous environmental factor and dominant susceptibility locus in our simulations. In practice, the investigator may be interested in an alternative type of environmental factor (e.g., continuous, multinomial) or genetic model (e.g., additive, codominant). Both steps in our method can naturally be extended to accommodate any parameterization of the environmental exposure or genetic coding. The absolute power of these extensions would depend on the underlying data distributions, but we would expect similar increases in power for our 2-step approach relative to a 1-step approach.

A typical GWAS is conducted on a large sample size to achieve power to detect modest-sized effects at genome-wide significance after correction for multiple testing. We considered a scenario with only 500 cases and controls genotyped on 10,000 markers for our base model. With an increase in sample size, power to detect a marker involved in a $G \times E$ interaction would increase for both the 1- and 2-step methods. An increase in number of markers could increase or decrease power, depending on whether the increase in linkage disequilibrium between the disease susceptibility locus and the markers offsets the penalty for a larger number of tests. However, we would expect variations in sample size and number of markers to affect both the 1- and 2-step approaches similarly and therefore not affect the relative comparison of power for the 2 methods.

We simulated scenarios in which a proportion (p_{ge}) of the available null markers were associated with the environmental factor in the population to establish that our 2-step approach preserves the desired type I error rate. It is possible for a causal locus to influence disease risk through an interaction with some environmental factor and simultaneously to be associated with the same environmental exposure in the

Table 2. Percentage of Replicates With a Ranked P Value for Disease Susceptibility Locus in the Top 10 or Top 25 Single Nucleotide Polymorphisms for 1-Step and 2-Step Tests for Gene-Environment Interaction

Model ^a	Top 10		Top 25	
	1 Step	2 Step	1 Step	2 Step
Base ^b	0.79	0.94	0.86	0.97
Disease susceptibility locus allele frequency (q_A)				
0.1	0.57	0.85	0.68	0.91
0.3	0.78	0.93	0.86	0.95
Exposure prevalence (p_E)				
0.1	0.26	0.51	0.35	0.63
0.25	0.66	0.89	0.76	0.94
Effect sizes (R_g , R_e , R_{ge})				
123	0.71	0.91	0.81	0.95
213	0.76	0.93	0.84	0.96
223	0.68	0.91	0.79	0.94
Population gene-environment association (p_{ge})				
0.01	0.76	0.94	0.85	0.97
0.05	0.79	0.91	0.86	0.94
0.30	0.75	0.84	0.84	0.91
0.95	0.75	0.76	0.84	0.85
No. of markers (M)				
25,000	0.69	0.93	0.78	0.95
50,000	0.65	0.87	0.74	0.92
Step 1 threshold (α_1)				
0.01	0.81	0.92	0.88	0.93
0.1	0.76	0.91	0.85	0.95

^a Variations to model specification refer to a change to a parameter setting in the "base" model. All other parameters remain constant.

^b $M = 10,000$, number of cases/controls = 500/500, $q_A = 0.2$, $p_e = 0.5$, $R_g = 1$, $R_e = 1$, $R_{ge} = 3$, $p_{ge} = 0$, $\alpha_1 = 0.05$.

general population. Under this scenario, the population-level association between the causal locus and the environmental factor would affect the power of our screening step. For example, a population-level $G \times E$ association at the causal locus in the opposite direction of the $G \times E$ interaction effect may reduce the power of our method. On the other hand, a positive $G \times E$ interaction combined with a positive $G \times E$ association in the population would inflate the estimated $G \times E$ association in our screening step and would likely increase the overall power of our 2-step test.

Incorporating a screening step to improve power in genetic analysis has been proposed in other contexts. For example, Van Steen et al. (10) developed a 2-step method for genome-wide association tests of genetic main effects using family data. Their screening step was based on a regression model using between-family information, and it was statistically independent of the family-based association test used in the second analysis step. Similar to their approach, our proposed analysis begins with a potentially biased test in the

first step that is designed to efficiently screen for potentially important SNPs and then uses an unbiased second step to ensure an overall valid procedure. Millstein et al. (11) also used a screening step in their Focused Interaction Testing Framework software to identify genes involved in $G \times G$ interactions in a study of many candidate loci. Although this screening test was biased in the presence of population-level association among genes, their second-step model ensured that the overall Focused Interaction Testing Framework approach was unbiased. Our results, in combination with those of Van Steen et al. and Millstein et al., demonstrate that well-designed 2-step approaches can lead to improved power in a wide range of genetic applications.

The additional power of our 2-step procedure comes from exploiting independent information provided by oversampling of cases relative to their prevalence in the population. In the presence of $G \times E$ interaction, this oversampling of cases induces an association between G and E in the combined case-control sample. Although it would be possible to develop an alternative 1-step test based on a likelihood that incorporates this additional information, such a test would not preserve the type I error in the presence of population-level $G \times E$ association. In the GWAS context, however, we can use the additional information derived from the oversampling of cases in a screening step to reduce the number of SNPs to be tested in the second step. When the power of the first step test is high, the chance that a true positive will be carried to the second step is also high. At the same time, a large number of null SNPs will be eliminated by the first-step screen, which reduces the multiple testing burden and results in our observed, overall gain in power to detect interaction at the causal locus.

Our 2-step approach is preferable to a case-only analysis of affected individuals, since the latter will have an inflated type I error rate in the presence of population $G \times E$ association. Even if a small subset of null SNPs has a population $G \times E$ association, one would expect several thousand false-positive results using a case-only analysis of interaction given the overall number of SNPs being tested. On the other hand, the type I error of our 2-step procedure is maintained because the second-step test is unbiased and the 2 steps are independent. Thus, even if there is a strong association in the population between E and a specific null SNP such that the SNP passes our first-step screen, the type I error rate for our overall 2-step procedure will be maintained.

Kraft et al. (12) proposed a powerful 2-df test for assessing genetic main effects and interactions jointly. They showed that under a wide variety of parameter settings, the 2-df test was often more powerful than a test of the main effect or the traditional test for $G \times E$ interaction. Although it is possible that a 2-df test would be more powerful than our 2-step testing framework, many investigators conducting a GWAS will want to begin with a full scan for genetic main effects. After conducting such a scan, use of the 2-df test would assess redundant information through partial retesting of the main effect. If a secondary goal is to detect genes involved in $G \times E$ interactions, our method allows it to be performed independently of the main-effect scan.

Several reported genome-wide association studies have conducted an initial scan of all available genotypes for main

effects by ignoring heterogeneity between exposure classifications (13–17). It is possible that by focusing completely on main effects, SNPs with disease associations modified by some environmental factor were not detected. Still, it has been argued that focusing on $G \times E$ interaction is not advantageous over testing for main effects (18). However, if there is strong evidence that an environmental factor contributes to risk and possibly modifies a genetic effect, it is potentially important to define a testing strategy that uses independent information in a second scan across the available markers. We developed this method in order to use this additional information to detect genetic heterogeneity across subgroups that might otherwise be missed in a direct main effect test. Although we focused on subgroups defined by an environmental factor, our approach can be used to assess heterogeneity across racial/ethnic groups, genotypes at another locus ($G \times G$ interaction), or any other variable that can be used to classify study subjects.

Our method relies on a priori knowledge of factors that might be expected to modify the risk of genotype on disease. For example, the Children's Health Study, a prospective cohort study designed to investigate respiratory outcomes in children in 12 communities throughout southern California, has shown evidence to suggest that both regional air quality and proximity to traffic contribute to risk of asthma, reduced lung function growth, and other respiratory outcomes (19–23). For the GWAS being conducted in this cohort, simply a scan of the main effects that ignored the potential modification of genetic effects by air pollutants might lead investigators to miss genetic variants that are important determinants of complex respiratory diseases.

We have shown in the context of a GWAS that utilizing ascertainment information through a screening step of available markers can lead to substantial increases in power to detect a gene involved in a $G \times E$ interaction. We furthermore showed that this 2-step approach is more robust to changes in environmental exposure, minor allele frequency, and so forth, than the traditional 1-step test for identifying highly ranked SNPs. Our approach therefore has the potential to increase the yield of a given GWAS by identifying additional, important loci that act in concert with an environmental or other factor to influence risk of a complex disease.

ACKNOWLEDGMENTS

Authors affiliation: Department of Preventive Medicine, University of Southern California, Los Angeles, California (Cassandra E. Murcay, Juan Pablo Lewinger, W. James Gauderman).

This work was supported in part by the National Institute of Environmental Health Sciences (T32 ES013678, 5P30ES007048, 5P01ES009581, R826708, RD831861, 5P01ES011627, 5R01ES014447, and 5R01ES014708), the National Heart, Lung, and Blood Institute (5R01HL087680, 5R01HL61768, and 5R01HL76647), the National Human Genome Research Institute (P50 HG 002790-02), and the Hastings Foundation.

Conflict of interest: none declared.

REFERENCES

1. Ito H, Matsuo K, Hamajima N, et al. Gene-environment interactions between the smoking habit and polymorphisms in the DNA repair genes, APE1 Asp148Glu and XRCC1 Arg399Gln, in Japanese lung cancer risk. *Carcinogenesis*. 2004;25(8):1395–1401.
2. Stern MC, Johnson LR, Bell DA, et al. XPD codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2002;11(10 pt 1):1004–1011.
3. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996;273(5281):1516–1517.
4. Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet Epidemiol*. 2007;31(5):365–375.
5. Zhao J, Jin L, Xiong M. Nonlinear tests for genomewide association studies. *Genetics*. 2006;174(3):1529–1538.
6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc (B)*. 1995;57:289–300.
7. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol*. 1996;144(3):207–213.
8. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol*. 2002;155(5):478–484.
9. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med*. 1994;13(2):153–162.
10. Van Steen K, McQueen MB, Herbert A, et al. Genomic screening and replication using the same data set in family-based association testing. *Nat Genet*. 2005;37(7):683–691.
11. Millstein J, Conti DV, Gilliland FD, et al. A testing framework for identifying susceptibility genes in the presence of epistasis. *Am J Hum Genet*. 2006;78(1):15–27.
12. Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered*. 2007;63(2):111–119.
13. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*. 2007;447(7148):1087–1093.
14. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet*. 2007;39(7):870–874.
15. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena R, Voight BF, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*. 2007;316(5829):1331–1336.
16. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*. 2007;316(5829):1341–1345.
17. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*. 2007;316(5829):1336–1341.
18. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*. 2001;358(9290):1356–1360.

19. Gauderman WJ, Avol E, Gilliland F, et al. The effect of air pollution on lung development from 10 to 18 years of age. *N Engl J Med*. 2004;351(11):1057–1067.
20. Gauderman WJ, Avol E, Lurmann F, et al. Childhood asthma and exposure to traffic and nitrogen dioxide. *Epidemiology*. 2005;16(6):737–743.
21. Gauderman WJ, Vora H, McConnell R, et al. Effect of exposure to traffic on lung development from 10 to 18 years of age: a cohort study. *Lancet*. 2007;369(9561):571–577.
22. McConnell R, Berhane K, Gilliland F, et al. Air pollution and bronchitic symptoms in Southern California children with asthma. *Environ Health Perspect*. 1999;107(9):757–760.
23. McConnell R, Berhane K, Gilliland F, et al. Asthma in exercising children exposed to ozone: a cohort study. *Lancet*. 2002;359(9304):386–391.
24. van der Vaart AW. *Asymptotic Statistics*. Cambridge, United Kingdom: Cambridge University Press; 1998.

APPENDIX 1

The Step-1 and Step-2 Test Statistics Are Asymptotically Uncorrelated

We assume a binary exposure E and binary genotype G , although the following result can be readily extended to multilevel exposures and genotypes. Let us consider the $2 \times 2 \times 2$ table for a case-control study in which, as before, D is the binary case-control indicator:

	$D = 1$			$D = 0$	
	$G = 1$	$G = 0$		$G = 1$	$G = 0$
$E = 1$	n_{11}	n_{12}	$E = 1$	n_{01}	n_{02}
$E = 0$	n_{13}	n_{14}	$E = 0$	n_{03}	n_{04}

The standard interaction $G \times E$ odds ratio is $OR_2 = (n_{11}n_{14}/n_{12}n_{13})/(n_{01}n_{04}/n_{02}n_{03})$, and the $G \times E$ odds ratio pooling cases and controls is $OR_1 = (n_{11} + n_{01})(n_{14} + n_{04})/(n_{12} + n_{02})(n_{13} + n_{03})$. $\log(OR_1)$ is the numerator of a Wald test statistic for step 1, and $\log(OR_2)$ is the numerator of a Wald statistic for step 2. We show that the asymptotic covariance $\text{Cov}(\log(OR_1), \log(OR_2)) = 0$ by using the delta method. The delta method establishes that if a random vector \mathbf{T}_n is asymptotically multivariate normal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as $n \rightarrow \infty$, then a differentiable transformation $f(\mathbf{T}_n)$ is asymptotically multivariate normal $N(f(\boldsymbol{\mu}), Df(\boldsymbol{\mu})^T \boldsymbol{\Sigma} Df(\boldsymbol{\mu}))$, where Df is the matrix of first-order derivatives of f (24).

The vectors $\mathbf{Y} = (n_{11}, n_{12}, n_{13}, n_{14})$ and $\mathbf{Z} = (n_{01}, n_{02}, n_{03}, n_{04})$ of observed counts in cases and controls are independent and have a multinomial distribution $\text{Mult}(p_1, p_2, p_3, p_4, n_1)$ and $\text{Mult}(q_1, q_2, q_3, q_4, n_0)$, respectively, where n_1 is the number of cases; n_0 is the number of controls; p_1, p_2, p_3, p_4 are the cell frequencies of the 2×2 $G \times E$ table for the cases; and q_1, q_2, q_3, q_4 are the cell frequencies of the 2×2 $G \times E$ table for the controls. By the standard approximation to the multinomial distribution, $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ is asymptotically normal with mean $\boldsymbol{\mu} = E[\mathbf{X}] = (n_1 p_1, n_1 p_2, n_1 p_3, n_1 p_4, n_0 q_1, n_0 q_2, n_0 q_3, n_0 q_4)$ and partitioned covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} n_1 p_1(1-p_1) & -n_1 p_1 p_2 & -n_1 p_1 p_3 & -n_1 p_1 p_4 & 0 & 0 & 0 & 0 \\ -n_1 p_1 p_2 & n_1 p_2(1-p_2) & -n_1 p_2 p_3 & -n_1 p_2 p_4 & 0 & 0 & 0 & 0 \\ -n_1 p_1 p_3 & -n_1 p_2 p_3 & n_1 p_3(1-p_3) & -n_1 p_3 p_4 & 0 & 0 & 0 & 0 \\ -n_1 p_1 p_4 & -n_1 p_2 p_4 & -n_1 p_3 p_4 & n_1 p_4(1-p_4) & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & n_0 q_1(1-q_1) & -n_0 q_1 q_2 & -n_0 q_1 q_3 & -n_0 q_1 q_4 \\ 0 & 0 & 0 & 0 & -n_0 q_1 q_2 & n_0 q_2(1-q_2) & -n_0 q_2 q_3 & -n_0 q_2 q_4 \\ 0 & 0 & 0 & 0 & -n_0 q_1 q_3 & -n_0 q_2 q_3 & n_0 q_3(1-q_3) & -n_0 q_3 q_4 \\ 0 & 0 & 0 & 0 & -n_0 q_1 q_4 & -n_0 q_2 q_4 & -n_0 q_3 q_4 & n_0 q_4(1-q_4) \end{bmatrix}.$$

If $f(\mathbf{X}) = (\log(OR_2(\mathbf{X})), \log(OR_1(\mathbf{X})))$, we have

$$Df(\boldsymbol{\mu}^T) = \begin{bmatrix} \frac{1}{n_1 p_1} & \frac{-1}{n_1 p_2} & \frac{-1}{n_1 p_3} & \frac{1}{n_1 p_4} & \frac{-1}{n_0 q_1} & \frac{1}{n_0 q_2} & \frac{1}{n_0 q_3} & \frac{-1}{n_0 q_4} \\ \frac{1}{n_1 p_1 + n_0 q_1} & \frac{-1}{n_0 p_2 + n_0 q_2} & \frac{-1}{n_1 p_3 + n_0 q_3} & \frac{1}{n_1 p_4 + n_0 q_4} & \frac{-1}{n_1 p_1 + n_0 q_1} & \frac{-1}{n_0 p_2 + n_0 q_2} & \frac{-1}{n_1 p_3 + n_0 q_3} & \frac{-1}{n_1 p_4 + n_0 q_4} \end{bmatrix}$$

and by the delta method, the asymptotic covariance matrix of $f(\mathbf{X})$ is

$$Df(\mu)^T \Sigma Df(\mu) = \begin{bmatrix} \frac{1}{n_1 p_1} + \frac{1}{n_1 p_2} + \frac{1}{n_1 p_3} + \frac{1}{n_1 p_4} + \frac{1}{n_0 q_1} + \frac{1}{n_0 q_2} + \frac{1}{n_0 q_3} + \frac{1}{n_0 q_4} & 0 \\ 0 & g(p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4) \end{bmatrix},$$

where $g = (n_1, n_0, p_1, p_2, p_3, p_4, q_1, q_2, q_3, q_4)$ is a rather lengthy expression involving $n_1, n_0, p_1, p_2, p_3, p_4, q_1, q_2, q_3$, and q_4 . The asymptotic joint distribution of $\log(\text{OR}_1)$ and $\log(\text{OR}_2)$ is then bivariate normal with covariance given by the off-diagonal entry of the 2×2 matrix above, that is, zero. Note that this result holds for *any* values of the cell frequencies and thus for any values of the underlying model parameters. This establishes the asymptotic independence of the 2 Wald statistics and, because they are asymptotically equivalent, of the corresponding likelihood ratio test statistics that we propose.

APPENDIX 2

The 2-Step Test Preserves the Type I Error

Let M be the total number of SNP markers, \mathcal{M}_0 be the subset of SNPs that are true negatives, \mathcal{M}_1 be the subset of SNPs that are true positives (i.e., for which there is $G \times E$ interaction), and M_0 and M_1 be the number of SNPs in \mathcal{M}_0 and \mathcal{M}_1 , respectively. Let T_{1k}, T_{2k} be the first- and second-step likelihood ratio statistics for SNP k , $1 \leq k \leq M$, α the desired genome-wide type I error level, and $\alpha_1 \approx \Pr(T_{1k} > c_1 | H_0)$ the step-1 type I error.

By a standard Bonferroni inequality argument, the genome-wide type I error is guaranteed to be less than α if the critical values c_1 and c_2 are chosen so that $\Pr(T_{1k} > c_1, T_{2k} > c_2) \leq \alpha/M_0$ for $k \in \mathcal{M}_0$. Because T_{1k} and T_{2k} are (asymptotically) independent (Appendix 1), it is equivalent to requiring

$$\Pr(T_{1k} > c_1) \leq \frac{\alpha}{M_0 \times \Pr(T_{1k} > c_1)} \approx \frac{\alpha}{M_0 \alpha_1}. \quad (\text{A1})$$

Now, the number of true negative SNPs tested in the second step can be written as $m_0 = \sum_{k \in \mathcal{M}_0} I(T_{1k} > c_1) = \sum_{k \in \mathcal{M}_0} I_k$, where $I_k = I(T_{1k} > c_1)$ and $I(\cdot)$ is the indicator function. This is a sum of M_0 nonindependent (because of linkage disequilibrium) Bernoulli random variables with probability of success $\approx \alpha_1$. The expected number of true negative SNPs to be tested in step 2 is therefore $E[m_0] \approx M_0 \alpha_1$, which is the denominator on the right-hand side of equation A1. For small α_1 , $\text{Var}(m_0) = \text{Var}(\sum_{k \in \mathcal{M}_0} I_k)$ is small (relative to M_0) because $\text{Var}(I_k) \approx \alpha_1(1 - \alpha_1)$, $\text{Cov}(I_i, I_j) \approx 0$ for SNP pairs in low linkage disequilibrium with each other (the vast majority), and $|\text{Cov}(I_i, I_j)| \leq \alpha_1(1 - \alpha_1)$ for pairs of SNPs in linkage disequilibrium with each other. Therefore, m_0 is close to its expectation $E[m_0]$ with high probability, and the right-hand side of equation A1, $\alpha/M_0 \alpha_1 \approx \alpha/E[m_0]$, is in turn approximately α/m_0 . Since the total number of SNPs tested in the second step is $m \geq m_0$, it suffices to choose c_2 as the $1 - \alpha/m$ quantile of a chi-square distribution with 1 df to satisfy equation A1 or, equivalently, to reject the null hypothesis if the second-step P value is smaller than α/m .